DOCUMENT RESUME

ED 335 374 TM 016 976

AUTHOR Werner, Eric
TITLE Component vs. Composite Grading: Outcome Quality and Passing Rate as Functions of Number of Tests, Test Correlation, and Reliability.
PUB DATE Apr 91
NOTE 23p.; Paper presented at the Annual Meeting of the American Educational Research Association and the National Council on Measurement in Education (Chicago, IL, April 3-7, 1991).
PUB TYPE Reports - Research/Technical (143) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS Accrediting Agencies; Agency Role; Certification; Comparative Analysis; Computer Simulation; *Correlation; *Licensing Examinations (Professions); *Pass Fail Grading; Quality Control; *Scores; Scoring; Test Content; *Test Reliability; Test Use
IDENTIFIERS *Composite Scores; *Multiple Cutoff Scores

ABSTRACT

Credentialing agencies that use multi-test examinations (MTEs) should be concerned with the quality of pass/fail outcomes as well as with the proportion of candidates passing. This study addressed three questions: (1) the effects on passing rate of MTE reliability, number of tests comprising a MTE, and correlations between pairs of component tests; (2) the effects on outcome quality of test reliability and correlations between pairs of component tests; and (3) how component (multiple-cutoff) and composite (compensatory) grading differ in terms of outcome quality when evaluated against the same standard. Using three independent variables (the number of tests comprising a MTE, the tests' reliabilities, and their true-score correlations) that define 18 conditions, large-sample data sets (1,000 scores) were computer-generated and analyzed under classical true-error conditions. Results show that: chances of candidate success on component-graded MTEs varied dramatically with the technical characteristics of the MTE; for component-graded and composite-graded MTEs, the quality of pass/fail outcomes varied significantly according to test characteristics; and the two grading strategies produced different distributions of candidates across outcome-quality categories. Implications for licensing MTE construction are discussed. Four tables present study data. An attachment illustrates the formula for controlling correlations among the distributions. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

Eric Werner

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

ED 374 784

COMPONENT vs. COMPOSITE GRADING: OUTCOME QUALITY AND PASSING RATE AS FUNCTIONS OF NUMBER OF TESTS, TEST CORRELATION, AND RELIABILITY

Eric Werner
Colorado Department of Regulatory Agencies

TM016976

BEST COPY AVAILABLE

Component vs. Composite Grading: Outcome Quality and Passing Rate as Functions of Number of Tests, Test Correlation, and Reliability[1]

Eric Werner
Colorado Department of Regulatory Agencies

Credentialing agencies that use multi-test examinations should be concerned with the quality of pass/fail outcomes as well as with the proportion of candidates passing. This study addressed three questions: (1) What are the separate and combined effects on examination passing rate of test reliability, number of tests composing an examination, and correlations between pairs of component tests? (2) What are the separate and combined effects on outcome quality of test reliability and correlations between pairs of component tests? (3) How do component (multiple-cutoff) and composite (compensatory) grading differ in respect to outcome quality when evaluated against the same standard? To explore these questions, large-sample data sets are computer-generated and analyzed under classical true-error conditions. Results demonstrate that (1) the chances of candidate success on component-graded examinations varies dramatically with the technical characteristics of the examination; (2) the quality of pass/fail outcomes of both component- and composite-graded examinations varies significantly according to test characteristics; and (3) largely due to typical pass rate differences between the two grading strategies, they produce differing distributions of candidates across outcome-quality categories. Implications for the development of credentialing examinations are discussed.

Credentialing organizations often base accept/reject determinations on the results of examinations consisting of two or more multiple-choice tests.. Some require that candidates satisfy a passing standard for each component test, whereas others sum or average candidate component scores and then require satisfaction of a single passing standard for the resulting composite. The first approach may be termed *component grading*, and the second *composite grading*. Regardless of which approach is employed, the use of multiple tests has significant implications for which and how many candidates are credentialed. This paper is concerned with how decision quality and passing rate are affected by choice of grading strategy and by several factors that characterize multi-test examinations. The factors of interest are test reliability, number

---

[1] Paper presented at the annual meeting of the American Educational Research Association and the National Council on Measurement in Education, April, 1991 in Chicago, Illinois.

of tests composing an examination, and the correlations among the tests that are used.

Credentialing programs aim to accept only candidates who are sufficiently prepared to practice competently and to reject only those who are not. Therefore, program officials are interested in the quality of their test-based accept/reject decisions. They are interested also in their examination passing rates. Even though the criterion-referenced perspective appropriately characterizes many, if not most, programs, passing rate continues to be an issue of public and professional interest. Examination programs having low passing rates may be suspect of obstructing the professional aspirations of qualified persons, and programs with high passing rates may be regarded as ineffective standards of competency assurance.

In theory, the grading strategy that an agency chooses should reflect its competency objectives. Component grading reflects the aim of assuring that every candidate is acceptably prepared with respect to every part of an examination. Composite grading allows for interaction among the component tests of an examination so that strong performance on one test can offset weak performance on another. In practice, however, competency objectives are seldom articulated precisely and the choice between strategies usually reflects a variety of technical, professional, and administrative factors. As a result, some organizations use a component grading approach even though they almost certainly could not demonstrate the necessity for independent standards for every component. Others organizations use composite grading when the nature of the subjects tested and their relationships to competent professional practice would justify component grading.

There are probably many circumstances in which a sharp distinction between component and composite grading is neither made nor appropriate. For example, a licensing board might use composite grading with the expectation that persons who pass the examination do not perform poorly on any of its components. From this perspective, there should be limited compensation among the parts, a reasonable perspective so long as each part is carefully defined, reliable, and distinctly important to professional practice. However, the minimal acceptable preparedness of successful candidates on all examination parts is almost always assumed rather than assured when pure composite grading is used.

Whenever decisions are based on test performance, a distinction must be made between the consistency and the accuracy of those decisions. Consistency concerns the similarity of outcomes over independently made assessments. However, since an examination might misclassify persons consistently, it is important to evaluate the accuracy of test-based decisions. Accuracy is the extent to which examination outcomes based on observed scores are the same as those that would be made on the basis of true scores. Accuracy,

rather than consistency, is relevant to this paper's focus on outcome quality.

Accuracy ordinarily is evaluated in relation to four kinds of decision outcomes. Expressed as four mutually exclusive candidate categories resulting from assessment by a single-test, these are:

1. true positives--candidates whose true scores and observed scores satisfy the standards applied to them (Equivalency and comparability of true and observed standards are implied as, for example, when both are expressed as the same percentage-correct value.)

2. false positives--candidates whose true scores are not up to standard but whose observed scores are

3. true negatives--candidates whose true scores and observed scores fail to satisfy the standards applied to them

4. false negatives--candidates whose true scores are up to standard but whose observed scores are not

Most accuracy research has involved efforts to estimate or minimize the probabilities of false positive and false negative outcomes (e.g., Wilcox, 1977). Some of this work has concerned the determination of optimal passing scores or suitable test lengths under conditions where false positive and false negative classification errors may or may not be equally serious (Hunyh, 1976, 1980; Novick & Lewis, 1974). Developing indices to measure the success of test-based decision making has been another subject of technical work (van der Linden & Mellenbergh, 1977). Much of this work on decision accuracy has involved very short, single-test examinations of narrowly defined educational objectives.

Algina and Gross (1979) directly addressed the subject of decision quality in the context of a multi-test examination where the distinction between component and composite grading was important. A six-part examination was administered to medical students at the end of their first year of study in order to determine whether they had mastered medical science course work. Concern had been expressed that a composite grading approach might allow students to advance to a three-year clinical curriculum despite deficiencies in particular subjects. Algina and Gross found that the concern was misplaced. Although approximately 20% of the successful students tested in the course of four years failed one or more subtests, well over 90% of these failed only one of the six subtests composing the examination.

Gross (1991) researched a 1989 administration of the Basic Science and the Clinical Science examinations of the National Board of Examiners in Optometry. Basic Science is composed of five sections and Clinical Science is composed of seven. The aim of the study

was to evaluate the effect of multiple criteria on pass/fail rates. However, its findings have implications for outcome quality as well. When passing scores were applied at the level of the entire Basic Science examination, 78% of the candidates passed. Imposition of section passing scores reduced the passing rate to between 50% and 5% (depending on the nature of the section passing scores used). In the case of Clinical Science, a total-examination passing score passed 84% of the candidates. Section passing scores resulted in passing rates between 39% and 2%. Gross emphasized that passing rate reductions associated with the creation of additional test sections under component grading conditions do not imply that higher quality credentialing outcomes are achieved. Instead, the reductions might be artifacts of the arbitrary erection of licensing hurdles. He also identified section length (and the closely related factor of reliability) and section interrelatedness as determinants of passing rate. The importance of reliability to outcome quality is supported by the findings of Livingston and Wingersky (1979).

The aforementioned categories of outcome accuracy are not directly applicable to the multi-test situation. Here one has more than one test coupled with a single pass/fail outcome that reflects a decision rule involving each test or a composite of all of them. This complicates things, especially in relation to composite grading. For example, a candidate who is successful under composite scoring might have, on at least one component test, both true and observed scores that are less than some applicable component-test performance standard. He or she is a true negative on each such component, but is something like a false positive in respect to the composite result.

Of course, the complication involving composite grading is created by taking the position that component true-score status is relevant to evaluating the quality of composite-grading outcomes. This position is pragmatic rather than absolute. It allows the comparison of component grading with composite grading in terms of the same criterion of true-score status. Such comparison is useful since most credentialing organizations that use composite grading probably would much prefer successful candidates who perform reasonably well on all or almost all components than those whose success is highly compensatory and involves, say, chance-level performance on some component tests.

To accommodate the multi-test situation, one can replace the false positive and false negative categories with, respectively, the categories of *successful-deficient* and *unsuccessful-sufficient*. A successful-deficient candidate is a success (under either component or composite grading), but has a true-score deficiency on one or more component tests. An unsuccessful-sufficient candidate is a failure (under either component or composite grading), but is true-score sufficient on all component tests. (Exact definitions of the two categories are given in the Analysis section of this paper.)

The probabilities of these kinds of error can be used to index assessment outcome quality.

The present study addressed three questions:

1. What are the separate and combined effects on examination passing rate of test reliability, number of tests composing and examination, and correlations between pairs of component tests? This question was studied in connection with component grading only. As explained later in this paper, the composite-grading passing rate was constant by design.

2. What are the separate and combined effects on outcome quality of test reliability and correlations between pairs of component tests?

3. How do component and composite grading differ in respect to outcome quality?

Joint distributions of true scores and observed scores are critical to evaluating outcome quality. These can be estimated (Livingston and Wingersky, 1979). Alternatively, one can simulate the necessary data sets within the constraints of a particular test score model. By generating both true and observed scores with desired characteristics, outcome quality can be studied under conditions that at least roughly approximate actual examinations. This was the approach taken in the present project.

The true and error scores generated satisfied four essential conditions of the classical true- and error-score model (Crocker and Algina, 1986):

1. The mean error score on any test was approximately zero.

2. The correlation between true and error scores was approximately zero.

3. The correlation of error scores over pairs of test was approximately zero.

4. The reliability of any test was equal to the ratio of true-to-observed variance and to the square of the correlation between true and observed scores.

Method

Independent Variables

The three independent variables of this research were the number of tests composing an examination, the reliabilities of these tests,

and their true-score correlations. Following are the levels of these variables:

number of component tests: 2, 3, 5

test reliability: .60, .80, .95

true-score correlation: .40, .70

Fully crossed, these variables define 18 conditions. Within a condition, all tests were to have the same reliability coefficient, and all pairwise true-score correlations were to be of a common value. Eighteen independent data sets were generated (as explained below) for these conditions. Both component and composite grading strategies were applied to each of these data sets.

Dependent Variables

Five dependent variables were used:

(1) proportion of candidates passing;

(2) probability of a successful-deficient outcome;

(3) probability of an unsuccessful-sufficient outcome;

(4) phi coefficient between candidate true status and candidate observed status); and

(5) number of successful-deficient records having substandard true scores on more than one test.

These variables are defined and discussed in the Analysis section of this paper.

Data Sets

For each of the *N* tests (*N*=2, 3, 5) within each of the 18 conditions, 1,000 random, standardized normal true scores were generated by computer. All pairwise correlations very closely approximated zero, departures representing negligible sampling error. *N* linear combinations each involving all *N* distributions within a condition were then defined such that the resulting *N* sets of true scores correlated as specified by the research design. Each combination assigned one of the *N* distributions the weight of one and the other(s) a common weight less than one. The formula for the nonunit weights is given in Attachment A.

Naturally, the means and standard deviations of the resulting true score distributions were affected by these transformations. Another linear transformation of each established a mean of 70 and

a standard deviation equal to the square root of 100 times the reliability coefficient specified for the condition.

For each test within a condition, 1,000 random, standardized normal error scores were computer-generated to have a mean of zero and a standard deviation equal to the square root of 100 times the difference between one and the reliability coefficient specified for the condition. Correlations between error score distributions very closely approximated zero as did cross correlations between true and error scores.

Normally distributed observed scores for each test were formed by adding the true and error parts for that test. The way in which true and error standard deviations had been set assured, in theory, that all observed-score standard deviations would be equal to 10. However, the fact that true-error correlations were extremely close to but not exactly equal to zero (due to sampling) resulted in nonzero covariance terms in the variance of observed scores. For purposes of subsequent computations, the standard deviations in question were linearly transformed so that they were precisely equal to 10. All observed score means were equal to 70.

These final transformations had no effect on the relative magnitudes of the true- and error-variance components in the observed-score distributions. For example, in the case of a test that was to have a reliability of .80, its total variance of 100 was 80% true-score variance and 20% error-score variance.

With means of 70 and standard deviations of 10, the resulting observed-score distributions within each of the 18 conditions were interpreted as percentage-correct, observed-score data sets. Since the reliability of each of these distributions had been manipulated directly by control of the relative size of true and error components in its variance, the issue of the number of items in the simulated tests was sidestepped. However, the intent was to simulate actual score distributions representing multiple-choice tests of considerable length. Credentialing examinations administered to large numbers of candidates often result in similar distributions, except that many are somewhat negatively skewed. Since the computations to be made on the simulated data sets would not emphasize extreme scores, this departure from typical distributions probably had little significance for the project objective of illustrating basic facts and relationships.

For each of the 18 conditions, composite observed scores were formed by averaging all component observed score distributions (2, 3, or 5) associated with that condition. Therefore, these distributions were also normal with means equal to 70 and standard deviations that were functions of component observed-score standard deviations, observed-score correlations, and the number of components averaged.

Passing Scores

A passing score of 65% was applied to each component observed-score distribution. Although this value equals the mean minus one half a standard deviation, it need not be interpreted in norm-referenced terms and is at least fairly representative of the range within which most criterion-referenced passing score for crede .tialing tests are set. Apart from this feature, the choice of 65% was arbitrary.

Because the true status and the observed status of candidates were to be compared, a passing score for each true-score distribution was established also. These true-score critical values were set equal to the true-score mean (70 in all cases) minus one half a standard deviation. The magnitude of these standard deviations depended directly and only on component test reliability, as previously discussed. To have applied the value of 65 to each true-score distribution, as was done with observed score distributions, would have been to use different standards at the true and observed score levels--standards that varied from one reliability condition to another. Under the procedure followed, the 65% passing standard applied to each observed-score normal distribution was statistically equivalent to its true-score normal counterpart (i.e., both represented the same normal deviate).

Composite grading required that there be a passing score for the combination of all component tests within each of the 18 conditions. In each condition, this passing score was set at the mean of the combination (70) minus one half its standard deviation. These values fell within the range of 65.4 to 66.8. Further, in terms of normal curve areas above and below these values, they were equivalent to component-test passing scores.

Because all true and observed distributions were normal, the passing scores one half of a standard deviation below the means assured that, within each condition, each of these distributions would have approximately 69% passing records. Of course, the passing rates resulting from component grading across these distributions would vary. Composite observed score distributions would in all cases have about 69% passing records. This matter is discussed further in the Results section.

Analyses

For each of the 18 conditions, pass/fail status on each test was determined at the true-score level and at the observed-score level. For purposes of the component grading strategy, a record was considered a success if every component observed score equalled or exceeded 65. It was a failure otherwise. For composite grading, a record was considered a success if the mean of its component observed scores equaled or exceeded the composite observed-score

passing standard for the condition with which it was associated. It was a failure otherwise.

If a record was a success under component grading as described above, but had at least one component true score that was less than the applicable true-score passing standard, it was considered a successful-deficient. If a record was a success under composite grading as described above, but had at least one component true score that was less than the applicable true-score passing standard, it was considered a successful-deficient. Thus, the criterion for a successful record being a deficient was the same for the two grading strategies investigated.

If a record was a failure under component grading, but had all component true scores at least equal to the applicable true-score passing standard, it was considered an unsuccessful-sufficient. If a record was a failure under composite grading, but had all component true scores at least equal to the applicable true-score passing standard, it was considered an unsuccessful-sufficient. As with successes, the criterion for an unsuccessful record being a sufficient was the same for the two grading strategies.

For each grading strategy as applied within each of the 18 conditions, the proportions of successful-deficient and of unsuccessful-sufficient outcomes were calculated as dependent variables. In each case, the base for successful-deficients was the number of successful records. The base for unsuccessful-sufficients was the number of unsuccessful records.

The more tests of which an examination is composed, the more likely it is that any passing record would have at least one true score less than the applicable standard, and the less likely it is that a failing record would have component true scores each of which satisfied the applicable standard. Thus, the successful-deficient and the unsuccessful-sufficient variables are confounded with the independent variable of number of tests. This independent variable was introduced to document its effects on passing rate, not outcome quality. Main effects of reliability and of true-score correlation will be computed across all levels of the number-of-tests variable. Since for each level of reliability and for each level of true-score correlation, there is equal representation within each level of number of tests, the confounding is irrelevant.

A phi coefficient was derived from these primary measures and used to evaluate the relationship between the true status and the observed status of records under each condition. If all component true scores of a record equaled or exceeded the applicable true-score passing standard for a condition, the value of one was assigned to the record. The value zero was assigned otherwise. In relation to component grading, the value of one was assigned to each record for which observed scores satisfied the component-grading passing rule defined previously; otherwise a zero was

assigned. For composite grading, the same one-zero assignment was used in connection with the passing rule for that strategy. Thus, true status could be correlated with observed status, separately for component grading and for composite grading. The higher such a correlation, the greater the accuracy of pass/fail determinations. By its definition, phi would reflect directly both the successful-deficient and unsuccessful-sufficient probabilities described above. It would, therefore, serve as an overall, summary measure of accuracy.

Livingston and Wingersky (1979) interpret the *G index* of Holley and Guilford (1964) as an index of efficiency of pass/fail decision making. The interpretation is based on the assumption that the seriousness of false-positive and false-negative decision errors does not depend on candidate true score. Since this assumption is consistent with the approach taken in the present project, G was computed and studied in relation to phi. G is derived directly from the proportion of all decisions that are either true-positives or true-negatives. G was developed in relation to the single-test situation. If one equates the single-test false-positive and false-negative categories with, respectively, the successful-deficient and unsuccessful-sufficient categories in the present multi-test study, G should be highly related to phi. In fact, the correlation between the two was .98. Had G been more independent of phi, its efficiency interpretation--though not entirely in harmony with the design of this project--might have been quite useful. However, given the size of its correlation with the more familiar phi, G was dropped from further consideration.

Another dependent variable was studied in relation to composite grading only. It was a count of the number of successful-deficient records having substandard true scores on more than just one test. In part because of its compensatory nature, composite grading ordinarily results in a higher passing rate than does component grading. Consequently, it is likely also to produce a higher probability of a successful-deficient error. Those who would advocate compensatory grading in a particular situation would be interested in the seriousness of the successful deficients that result. If, for example, almost all successful deficients have substandard true scores on one component only, perhaps a higher successful-deficient probability is an acceptable price to pay for other perceived advantages of compensatory grading.[2]

---

[2] Another approach would be to examine mean true scores for successful-deficients to determine the degree to which they are substandard. This method would be influenced significantly by the normal-curve model used in this study, true scores in the extreme lower tail of the normal distributions exerting a strong downward pull on the means computed. Because extreme scores are probably better represented with models other than the normal-curve, this method was not used in the present study.

Results

Comparing Component Grading With Composite Grading

Results reported in this section concern the individual and combined effects of independent variables on passing rates and outcome-quality measures. Results are reviewed separately for component and composite grading. However, the study was not designed to assure that scoring strategy alone distinguished these two conditions. They differed also in respect to passing rate--a constant of approximately 69% for composite grading and a variable rate (always less than 69%) for component grading. This simulates the reality of many credentialing situations in which the choice between composite and component grading involves consideration of major differences in the passing rates that would result under the alternatives.

Percentage-correct passing scores applied to credentialing tests usually fall within the range of 60% to 75%. Applying any such standard to each component test of an examination will ordinarily pass far fewer candidates than would be passed if the same value were applied only to the average of the component test scores. A decision on whether to use composite or component grading typically involves consideration of this fact. Therefore, it was reasonable to allow passing rate to vary between these two major conditions.

However, this project feature has implications for how results are interpreted. When component and composite grading differ in respect to an outcome-quality measure, should the difference be attributed to grading strategy per se or to passing rate? At least under the conditions of this project, the answer appears to be passing rate. Indeed, if the passing score for each composite condition were set such that the passing rate were the same as under the corresponding component condition, findings would be similar for the two conditions.

The primacy of passing rate rather than scoring strategy per se in determining the numerical values of accuracy measures and the effects of independent variables on these measures does not mean that comparing component grading with composite grading is of no practical interest or importance. The real-world use of a composite grading strategy almost always *does* involve a much higher passing rate than would the component-grading alternative and is sometimes justified precisely on this basis. Therefore, it is useful to compare the two strategies even as they differ in respect to passing rate--but without the implication that differences in decision accuracy or in independent variable effects are attributable to scoring method per se.

Relations Among Reliability, True-Score Correlation and Observed-Score Correlation

The design of the study controlled the level of true-score correlation among pairs of tests. These correlations did of course affect the correlations between observed scores on these tests, and the observed-score correlations did in turn affect dependent variables. However, test reliability also affected observed-score correlations by imposing an upper bound on these values. Thus, although reliability and true-score correlation are independent by design, the effects of both are in part manifest through the critical factor of observed-score correlation. The mean observed-score correlation over conditions where the true-score correlation was .70 was .55. For conditions in which the true-score correlation was .40, the mean observed-score correlation was .32.

Main Effects

Table 1 shows the separate effects of number of tests, reliability, and true-score correlation for the component-grading strategy. Table 2 shows the same kind of information for the composite-grading strategy.

TABLE 1
Dependent Variable Means As Functions of Independent Variables: Component Scoring

		Dependent Variable			
Independent Variable	Level	Pass Rate	S-D Prob	U-S Prob	Phi
No. tests	two	.54	--	--	--
	three	.44	--	--	--
	five	.32	--	--	--
Reliability	.60	.40	.18	.24	.58
	.80	.45	.14	.15	.71
	.95	.46	.07	.08	.85
True-score r	.40	.40	.16	.14	.70
	.70	.47	.10	.17	.73
Grand mean		.43	.13	.16	.71

Component grading

Number of Tests. As the number of tests composing an examination increased, the mean proportion of successful records decreased from well over half to less than a third. Since number of tests is not a variable of interest in connection with successful-deficient and unsuccessful-sufficient probabilities and since it is confounded with these dependent variables, its "effects" on them are not discussed or presented in Table 1.

Reliability. It was not expected that passing rate would be much changed by increased reliability. In fact, it gained six percentage points as reliability increased from .60 to .95. The relation between test reliability and the probability of a successful deficient was inverse. As reliability increased, this probability declined from nearly one in five to about one in fifteen. The effect of reliability on the probability of an unsuccessful sufficient was greater, but inverse also. Higher reliability reduced unsuccessful-sufficient odds from approximately one in four to about one in thirteen. The responsiveness of the two error probabilities to improved reliability was reflected in a sharp increase in the value of phi.

True-score correlation. True-score correlation had a modest effect on the dependent variables. Its increase resulted in a modest rise in passing rate and the mean unsuccessful-sufficient probability, and a small fall in the mean successful-deficient probability. Consequently, phi was hardly changed.

Component grading overall. The grand mean passing rate for component grading over all 18 conditions was .43. Successful-deficient and unsuccessful-sufficient outcomes were about equally likely, the respective grand means being .13 and .16. The grand mean phi coefficient was .71.

Composite grading

Number of tests. Since number of tests was of interest in relation to passing rate only, and passing rate under composite grading was a constant, no results are reported for this independent variable.

Reliability. Table 2 shows that reliability had a negligible effect on the mean successful-deficient probability, which declined only slightly as reliability increased. For each level of reliability, this mean probability was at least twice the corresponding probability under component scoring. The mean probability of a unsuccessful-sufficient was quite small under all reliability conditions, its value decreasing to nearly zero when reliability was at its maximum. The mean value of phi was moderately responsive to change in reliability. For reliability

of .60, phi was approximately equal to the corresponding value for component grading. As reliability increased, however, the phi for composite grading rose much less than did phi under component grading. The effect of reliability on phi for composite grading was less than half that for component scoring.

TABLE 2
Dependent Variable Means as Functions of Independent Variables: Composite Scoring

		Dependent Variable		
Independent Variable	Level	S-D Prob	U-S Prob	Phi
Reliability	.60	.35	.08	.59
	.80	.33	.04	.65
	.95	.32	.01	.71
True-score r	.40	.40	.04	.61
	.70	.27	.05	.70
Grand mean		.33	.05	.65

True-score correlations. As true-score correlation increased from .40 to .70, the mean probability of a successful deficient decreased substantially. The mean odds of an unsuccessful sufficient were very low under both true-score correlation conditions. Mean phi increased somewhat with increased correlation, reflecting the effect on successful deficients.

Composite grading overall. The grand mean successful deficient probability was .33, two and one half times the size of the corresponding grand mean for component grading. The grand mean probability for unsuccessful sufficients was at the other extreme, its value of .05 being less than a third the size of its component-grading counterpart. The grand mean phi was .65, somewhat less than its value under component grading.

Combinations of Independent Variables

For both component and composite grading, the effects of independent variables on dependent variables were additive. The effect of each independent variable was visually inspected within levels of each other independent variable, and all simple main effects were found to be in the same direction. Thus, from knowledge of main effects alone, one can readily identify the combinations of independent variables that will produce extreme

values of any dependent variable. Table 3 identifies extreme conditions for component grading, and Table 4 does so for composite grading. Because measures of outcome quality are confounded with the independent variable of number of tests, results shown for outcome quality variables are means taken over all levels of the independent variable.

Component grading

The difference between the maximum and minimum of each dependent variable was large. Passing rate was as low as .25 and as high as .58. Both error probabilities ranged from near zero to approximately .25. Phi reflected these probability variations with a minimum of only .56 and a maximum of .86.

TABLE 3
Maximal and Minimal Value Conditions of Dependent Variables: Component Grading

	Independent Variable			
Dependent Variable	No. tests	Reliability	True-score r	Max/Min
Pass Rate				
Maximum	two	.95	.70	.58
Minimum	five	.60	.40	.25
S-D Probability				
Maximum	--	.60	.40	.23
Minimum	--	.95	.70	.05
U-S Probability				
Maximum	--	.60	.70	.27
Minimum	--	.95	.40	.07
Phi				
Maximum	--	.95	.70	.86
Minimum	--	.60	.40	.56

Composite grading

The successful-deficient probability ranged from .26 to .42. Its minimum value was nearly equal to the maximum for component grading. Unsuccessful-sufficient odds ranged less and the maximum was much lower than even the minimum value for successful-deficient odds. This maximum was approximately equal to the minimum unsuccessful-sufficient probability for component grading. At its lower extreme, phi was about the same as under component grading, but its maximum was lower than that for component grading.

TABLE 4
Maximal and Minimal Value Conditions of Dependent Variables: Composite Grading

Dependent Variable	Independent Variable: No. tests	Reliability	True-score r	Max/Min
S-D Probability				
Maximum	--	.60	.40	.42
Minimum	--	.95	.70	.26
U-S Probability				
Maximum	--	.60	.70	.09
Minimum	--	.95	.40	.01
Phi				
Maximum	--	.95	.70	.76
Minimum	--	.60	.40	.54

Extent of True-Score Deficiencies For Composite Grading

Over all 18 composite-grading conditions, the mean proportion of successful-deficient records having substandard true scores on more than one test was .24. Reliability had a large effect on this variable. Overall, as reliability increased from .60 to .95, the mean proportion of successful-deficients with more than one substandard true score decreased from .30 to .18. Within the two-test condition, the dependent variable proportion declined from .25 to .00 as reliability rose. For three tests, the decline was from .31 to .13. For five tests, it was from .46 to .39.

Discussion

This study sought to demonstrate the effects of several test characteristics on the passing rate and pass/fail outcome quality of multi-test examinations. In credentialing, most such examinations are composed of multiple-choice tests. Findings discussed here hold for the parameters used in the current study and generalize to other circumstances to the degree that they are characterized by these same conditions.

Both component and composite grading were evaluated in terms of component true-score status. This may seem an unfair basis for evaluating composite grading, since the strategy is not intended to assure the sufficiency of all passing candidates on every component. However, it is reasonable for credentialing program administrators to express interest in the component-level preparedness of those who are passed and failed under a composite-grading standard. This project addressed that question by

evaluating the two grading strategies in terms of the same criterion.

The study's design created substantial passing-rate differences between component and composite grading. These differences, rather than a fundamental difference between the two strategies per se, explain most of the results found. However, the passing rates in this study are not too different from those encountered in actual application of the two strategies. For practical purposes, therefore, the two strategies are effectively different.

Study findings document and underscore the importance of test reliability to effective decision making. Under component grading, improving the reliability of tests can reduce successful-deficient and unsuccessful-sufficient outcomes from the range of about 20%-25% to nearly nothing. Further, this improvement in decision accuracy is associated with some gain in passing rate. With composite grading, where the probability of successful-deficient outcomes can be large, reliability improvement can very substantially reduce the proportion of passing candidates who have substandard true-score status on more than one component.

The correlation between true scores also has some noteworthy effects. For composite scoring in particular, its increase substantially lessens the probability of successful-deficients. Otherwise, its primary role is to affect passing rate under component scoring, lowering the rate as true-scores become more independent.

Comparison of component with composite grading shows the latter to be on average the considerably more risky strategy, if one places a high priority on avoiding successful-deficient outcomes. On the other hand, it offers negligible probability of unsuccessful-sufficients. Component grading presents a more balanced risk picture with fairly small and nearly identical odds of these two outcome categories. If one adds the two sets of odds for each grading strategy, the sums are quite similar.

The number of tests included in an examination graded under the component strategy greatly affects the proportion of candidates who pass. In this study, increasing the number from two to five decreases the proportion from nearly 55% to nearly 30%. In conjunction with low reliability and low component true-score correlations, the proportion falls further to 25%.

In view of the passing rate and outcome differences between component and composite grading, it is important that credentialing organizations carefully define their competency objectives and select the appropriate grading strategy. Justifying component grading on the thought that everyone benefits if all candidates are strong in all areas tested is too simplistic to support the relatively harsh pass/fail results of the strategy. Likewise,

establishing separate passing standards for each of several subjects simply because such practice is generally consistent with the way in which academic training is structured misses the point that credentialing examinations should be practice-related assessments, not extensions of classroom requirements. On the other hand, averaging or summing part scores just because it is convenient to do so, usually passes more candidates, and results in fewer complaints falls far short of what may be necessary if an examination is to protect the public against persons ill-prepared to practice safely and effectively.

If job analysis procedures were designed and implemented to better identify compensable and noncompensable knowledge areas, credentialing organizations would be able to base their competency objectives more solidly on empirical, practice-related information. More informed grading strategy decisions would result. A greater role for job analysis information probably also would promote more effective discussion between advocates of practice-based test design and those more inclined toward academically oriented test structure. It is the author's opinion that inappropriate use of component grading is more likely to be found where academically oriented knowledge classifications serve as the basis for test planning.

Probably most credentialing could be effective with composite grading or a mixture of composite and component grading. Even where two knowledge (or performance) areas differ distinctly from one another, it does not follow that component grading is needed. One must ask whether the difference is clearly significant to the public interest. Only an affirmative answer justifies pure component grading with its lower passing rate and lower probability of passing candidates who are true-score deficient in particular areas. If composite grading results in an unacceptable degree of compensatory test performance, a mixture of composite and component grading is likely to be a better remedy than is pure component grading.

References

Algina, J., & Gross, L. (1979). Decision-making in basic medical sciences examinations: Single versus multiple cutoff scores. *Evaluation and the Health Professions, 2,* 203-208.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* New York: Holt, Rinehart and Winston.

Gross, L. (1991). Examination standards for licensure: The reality of 75. *Journal of Optometric Education, 16,* 72-75.

Holley, J. W., & Guilford, J. P. (1964). A note on the G index of agreement. *Educational and Psychological Measurement, 24,* 749-753.

Hunyh, H. (1976). Statistical considerations on mastery scores. *Psychometrika, 41,* 65-78.

Hunyh, H. (1980). A nonrandomized minimax solution for passing scores in the binomial error model. *Psychometrika, 45,* 167-182.

Livingston, S. A., & Wingersky, M. S. (1979). Assessing the reliability of tests used to make pass/fail decisions. *Journal of Educational Measurement, 16,* 247-260.

Novick, M. R., & Lewis, C. (1974). Prescribing test length for criterion referenced measurement. In C. W. Harris, M. C. Alkin, & W. J. Popham (Eds.). *Problems in criterion-referenced measurement.* Los Angeles: Center for the Study of Evaluation, University of California.

van der Linden, W. J., & Mellenbergh, G. J. (1982). Optimal cutting scores using a linear loss function. *Applied Psychological Measurement, 19,* 593-599.

Wilcox, R. R. (1977). Estimating the likelihood of false positive and false negative decisions in mastery decisions: An empirical Bayes approach. *Journal of Educational Statistics, 2,* 289-307.

Attachment A: Formula for Controlling Correlations Among True-Score Distributions

Start with the formula for the correlation between two weighted composites, X and Y, each consisting of the same p components and differing only in terms of component weights. In X, each component except C_m carries weight k. C_m carries the weight of one. In Y, each component except C_n $(m \neq n)$ carries weight k. C_n carries the weight of one. There are N observations $(i=1, \ldots, N)$ on each variable.

$$r_{XY} = \frac{\frac{1}{N}\Sigma\left[(kC_{i1}+\cdots+C_{im}+\cdots+kC_{ip})(kC_{i1}+\cdots+C_{in}+\cdots+kC_{ip})\right]-\bar{X}\,\bar{Y}}{\sigma_X\,\sigma_Y}$$

The problem is to find k. Components are uncorrelated with one another and each is standardized to mean zero and variance one. The standard deviation of composite X equals the standard deviation of composite Y,

$$\sigma_x = \sigma_y = \left(\sigma_{C_m}^2 + k^2 \sum_{j=1}^{p} \sigma_{C_j}^2\right)^{1/2} = \left[1 + k^2(p-1)\right]^{1/2}, \quad j \neq m.$$

Because all component means are equal to zero, the mean of composite X equals the mean of composite Y equals zero. Therefore,

$$r_{XY} = \frac{\frac{1}{N}\Sigma\left[(kC_{i1}+\cdots+C_{im}+\cdots+kC_{ip})(kC_{i1}+\cdots+C_{in}+\cdots+kC_{ip})\right]}{1+k^2(p-1)}.$$

The numerator gives p^2 terms in five forms. Three forms (representing $p(p-1)$ terms) involve cross-products, and in each instance equal zero since all components are uncorrelated. One of the other forms is:

$$\frac{k}{N}\Sigma C_i^2 = k.$$

There are 2 such terms. The remaining form is:

$$\frac{k^2}{N}\Sigma C_i^2 = k^2.$$

There are $p-2$ such terms.

Therefore, the composite correlation formula can be written as

$$r_{XY} = \frac{(p-2)k^2 + 2k}{1+(p-1)k^2}.$$

This can be expressed in standard quadratic form as

$$[(p-1)r_{XY} - (p-2)]k^2 - 2k + r_{XY} = 0.$$

Solving for k gives

$$k = \frac{2 \pm \{4 - 4[(p-1)r_{XY} - (p-2)]r_{XY}\}^{1/2}}{2[(p-1)r_{XY} - (p-2)]}.$$

The smaller root is the weight needed.